PATENT ABSTRACTS OF JAPAN

(11)Publication number : 09-319767
(43)Date of publication of application : 12.12.1997

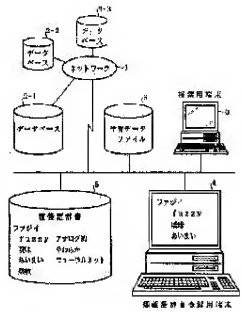
(51)Int.Cl. **G06F 17/30**

(21)Application number : 08-157722
(22)Date of filing : 29.05.1996

(71)Applicant : OKI ELECTRIC IND CO LTD
(72)Inventor : JIYOUFUU TOSHIHIKO

(54) SYNONYM DICTIONARY REGISTERING METHOD

(57)Abstract:
PROBLEM TO BE SOLVED: To generate a precise dictionary including newly coined words, etc., by experientially obtaining a key word which is highly possibly used at the same time of a key word to register from the present time from an actual retrieving result and selecting a synonym by setting this as reference, so as to widely extract words.
SOLUTION: When a data base is retrieved, various key words are inputted by OR combination or AND combination from a retrieving terminal 3. At this time, data on what kind of key word is combined by OR or AND is stored in a learning data file 6. At the time of registering a synonym through the use of a terminal for registering a synonym dictionary 4, a word which is highly possibly retrieved at the same time of some key word and a word of high similarity are fetched by utilizing the file 6 and these are displayed on a display. A registering person retrieves a proper key word while looking at the list and registers it to the synonym dictionary.



* NOTICES *

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

CLAIMS

[Claim(s)]
[Claim 1]In a case where a word with a near distance is semantically registered into a synonym dictionary for extending a keyword to a keyword for database retrieval, When making coincidence probability of a keyword of a lot into a standard showing nearness of a semantic distance, Inside of a group of a keyword inputted for a user's search of said database, As opposed to arbitrary keywords which amended coincidence probability of each keyword used OR combination having been carried out by study so that it might be made to increase at every input of the, and were made into a dictionary registration object, A synonym dictionary registration method displaying other keyword lists with said high coincidence probability, and registering other keywords selected from these lists into a synonym dictionary as a synonym to a keyword made into a dictionary registration object.
[Claim 2]In a case where a word with a near distance is semantically registered into a synonym dictionary for extending a keyword to a keyword for database retrieval, When considering it as a standard showing height of probability of appearing connection probability of a keyword of a lot simultaneously, Inside of a group of a keyword inputted for a user's search of said database, As opposed to arbitrary keywords which amended connection probability of each keyword used AND combination having been carried out by study so that it might be made to increase at every input of the, and were made into a dictionary registration object, A synonym dictionary registration method displaying other keyword lists with said high connection probability, and registering other keywords selected from these lists into a synonym dictionary as a synonym to a keyword made into a dictionary registration object.
[Claim 3]In a case where a word with a near distance is semantically registered into a synonym dictionary for extending a keyword to a keyword for database retrieval, As opposed to arbitrary keywords which memorized an erroneous input keyword inputted for a user's search of said database, and were made into a dictionary registration object, Extract a high thing of notation similarity out of said erroneous input keyword, and other keyword lists with high notation similarity are displayed, A synonym dictionary registration method registering other keywords selected from these lists into a synonym dictionary as a synonym to a keyword made into a dictionary registration object.

[Translation done.]

* NOTICES *

DETAILED DESCRIPTION

[Detailed Description of the Invention]
[0001]
[Field of the Invention]This invention relates to the synonym dictionary registration method used when searching the electronized document using a keyword.
[0002]
[Description of the Prior Art]For example, in trying to perform information retrieval on a network like the Internet, a huge quantity of a document serves as a retrieval object. These documents are drawn up in arbitrary languages by arbitrary thought from the first, and, moreover, contain many spelling errors, new words, etc. Therefore, when searching these using a keyword, in order to obtain suitable search results, a user needs various kinds of devices. For example, when searching using a fixed keyword, a meaning registers the well alike word into the keyword, and these words are simultaneously used for search. This method is called extension of a keyword. The method of drawing up the synonym dictionary for extending such a keyword automatically, For example, it is indicated in the following literature (Institute of Electronics, Information and Communication Engineers .TECHNICAL REPORT OF IEICE A.I. Artificial Intelligence95-24 (1995-09) PP15-22. "distributed news search system using a statistical thesaurus").
[0003]
[Problem(s) to be Solved by the Invention]By the way, there were the following issues which should be solved in the above conventional synonym dictionary registration methods. Before performing a search, the keyword and this which are beforehand used for search, and the near language of a meaning are listed and registered into a synonym dictionary. The registrant who takes charge of dictionary creation classified the keyword into parts of speech, such as a noun, a verb, and an adjective, beforehand, classified the keyword from still more various viewpoints, and has registered the word equivalent to the generic concept of each keyword, and a subordinate concept as a synonym. For example, they are given to the keyword "industry" by classifications, such as human activities, pneuma, and an act, and as a synonym, It sees from the same viewpoint, sees from "industry", "business", "a performance", and the viewpoint as a result, sees from the viewpoint as "production", "production increase", "decrease in production", and a process, and language, such as "foundation", "a classification", and "a provincial tour", is registered.
[0004]When it is going to search the literature about fuzzy logic, as a search condition, It is preferred to use language, such as adaptability, "analog", softness, and a "neural network", as a synonym in addition to "fuzzy" OR"FUZZY"OR "unreliable" OR ambiguity and this. Therefore, such language is registered as a synonym. However, it is not necessarily easy to choose a synonym widely and to register it in this way. A user's demand cannot be met when it searches using an insufficient synonym dictionary. Therefore, there was a problem that a big burden was placed on the registrant of a synonym dictionary.
[0005]Suitable search cannot be carried out if the keyword inputted in the case of search has a spelling error in at least 1 character. When it searches using techniques, such as match partial, in consideration of the difference in a prefix or the ending, a noise increases and it is not sometimes practical. Especially the thing that the synonym of sufficient quantity corresponding to this is set to the suitable keyword for the field by which a new technical term is produced every day, and is registered beforehand is dramatically difficult. Therefore, synonym dictionary registering operation is made simpler and construction of the system by which efficient high-precision search results

are obtained is desired.
[0006]
[Means for Solving the Problem]This invention adopts the next composition in order to solve the above point.
<Composition 1> In a case where a word with a near distance is semantically registered into a synonym dictionary for extending a keyword to a keyword for database retrieval, When making coincidence probability of a keyword of a lot into a standard showing nearness of a semantic distance, Inside of a group of a keyword inputted for a user's search of the above-mentioned database, As opposed to arbitrary keywords which amended coincidence probability of each keyword used OR combination having been carried out by study so that it might be made to increase at every input of the, and were made into a dictionary registration object, A synonym dictionary registration method displaying other keyword lists with the above-mentioned high coincidence probability, and registering other keywords selected from these lists into a synonym dictionary as a synonym to a keyword made into a dictionary registration object.
[0007]<Explanation> If some keywords are registered into a synonym dictionary as a synonym to a certain keyword, a keyword which a user gave will be automatically extended in the case of database retrieval, and retrieval precision will be raised at it. Even when a keyword of a lot is separate from coincidence probability in arbitrary documents contained in a database, it is the probability used by both somewhere. A keyword inputted by OR combination in the case of database retrieval with a actual user is mutually accepted that a semantic distance is near. Then, it is made to learn so that coincidence probability may be enlarged, whenever a group of the keyword is inputted. In this way, data used as the foundation for synonym dictionary registration is stored automatically. Since practical similarity is considered by this, difficulty of synonym dictionary registering operation is eased and accuracy at the time of carrying out fuzzy search of the database with techniques, such as a full-text search, is raised.
[0008]<Composition 2> In a case where a word with a near distance is semantically registered into a synonym dictionary for extending a keyword to a keyword for database retrieval, When considering it as a standard showing height of probability of appearing connection probability of a keyword of a lot simultaneously, Inside of a group of a keyword inputted for a user's search of the above-mentioned database, As opposed to arbitrary keywords which amended connection probability of each keyword used AND combination having been carried out by study so that it might be made to increase at every input of the, and were made into a dictionary registration object, A synonym dictionary registration method displaying other keyword lists with the above-mentioned high connection probability, and registering other keywords selected from these lists into a synonym dictionary as a synonym to a keyword made into a dictionary registration object.
[0009]<Explanation> Connection probability is the probability that a keyword of a lot will appear continuously simultaneously in arbitrary documents contained in a database. It is admited that a keyword inputted by AND combination in the case of database retrieval with a actual user has high probability of appearing simultaneously. Then, it is made to learn so that connection probability may be enlarged, whenever a group of the keyword is inputted. In this way, data used as the foundation for synonym dictionary registration is stored automatically. Since language with high probability that this will appear simultaneously semantically in addition to near language is also registered as a synonym, accuracy at the time of carrying out fuzzy search of the database with techniques, such as a full-text search, is raised.
[0010]<Composition 3> In a case where a word with a near distance is semantically registered into a synonym dictionary for extending a keyword to a keyword for database retrieval, As opposed to arbitrary keywords which memorized an erroneous input keyword inputted for a user's search of the above-mentioned database, and were made into a dictionary registration object, Extract a high thing of notation similarity out of the above-mentioned erroneous input keyword, and other keyword lists with high notation similarity are displayed, A synonym dictionary registration method registering other keywords selected from these lists into a synonym dictionary as a synonym to a keyword made into a dictionary registration object.
[0011]<Explanation> A mistake notational [many] is mixed with a keyword inputted in the case of database retrieval with a actual user. Then, if a large thing of notation similarity is taken out out of an actually used erroneous input keyword list and it registers with a synonym dictionary for a right keyword, it can search by correcting what is called a spelling error etc. automatically. Thereby,

difficulty of synonym dictionary registering operation is eased and accuracy at the time of carrying out fuzzy search of the database with techniques, such as a full-text search, is raised.
[0012]
[Embodiment of the Invention]Hereafter, an embodiment of the invention is described using an example.
<Example> Drawing 1 is a system block figure for operation of this invention. The method of this invention is enforced by a system as shown, for example in this figure. In the figure, arbitrary database 2-1,2-2 and 2-3 grade are connected to the network 1. It may be connected via linked another network and these databases can take various gestalten. This should be constituted by the Internet etc., for example.
[0013]Here, in order to carry out the full-text search of various kinds of articles on these databases using a keyword, a system as shown in this figure is prepared. First, the terminal 3 for search is established for search, and the synonym dictionary 5 is used for the keyword extension which should be searched. In order to generate and register this synonym dictionary 5, the terminal 4 for synonym dictionary registration and the study data file 6 used in this invention are formed. In this invention, when a keyword is actually inputted using the terminal 3 for search and various databases are searched, the keyword which were used is stored in the study data file 6. By this, the basic data for synonym dictionary registration is obtained.
[0014]When a registrant inputs the keyword used as a registering object, the terminal 4 for synonym dictionary registration reads the high word of the keyword and similarity from the study data file 6, and similarity displays it on high order partly. A registrant chooses a suitable thing out of these words, and registers with the synonym dictionary 5. By the example 1, use the study data file 6 as a coincidence probability learning data file, it is used as a connection probability learning data file by the example 2, and let them be addition, omission, and a substitution list file by the example 3. Although the above is an outline of this invention, the actual condition of the registration is hereafter explained using an example, respectively.
[0015]<Example 1> The explanatory view of the example 1 of operation is shown in drawing 2. As shown in this figure, the registrant 7 performs synonym registration processing via the registration interface 8. This registration interface 8 is included in the terminal 4 for synonym dictionary registration shown in drawing 1. Since the information which can be used for synonym registration of a keyword is extracted and saved from the search information 10, the coincidence probability learning data file 11 is formed. The similarity calculation module 9 is formed for similarity calculation. This similarity calculation module 9 shall also be included in the terminal 4 for synonym dictionary registration shown in drawing 1.
[0016]In actual search, when searching two or more keywords by OR combination, that which the meaning all bears a strong resemblance to dramatically mutually is contained in these keywords. Therefore, improvement in retrieval precision can be aimed at by registering these into the synonym dictionary 5. Mutually, although the keyword which makes carry out OR combination and is searched is not necessarily approaching somewhere in one document mutually, it is used simultaneously. Thus, the probability simultaneously used for one document is called coincidence probability. The keyword with this high coincidence probability can be mutually judged that similarity is high. This coincidence probability itself can be taken out by analyzing by carrying out the full-text search of each document which constitutes a database, for example. In this example 1, the degree for which a retrieving person uses and uses OR combination in the case of actual search amends such coincidence probability. Therefore, coincidence probability increases gradually, so that the degree searched with OR combination is high. Thus, in order to raise coincidence probability at every search, the coincidence probability learning data file 11 was formed.
[0017]If the method of this invention is explained in order according to this drawing 2, the search information 10 which the retrieving person inputted in the usual database retrieval will be first accumulated in the coincidence probability learning data file 11 via the similarity calculation module 9 in Step S1 (Step S2). This coincidence probability is calculated as follows.
Coincidence probability Wij = document number in which one of the number of times / word "i" with which the word "i", and the word "j" coincided, and the words "j" appeared -- (1)
[0018]Regarding similarity with some keywords Ki and Kj as Wij is known from the former. For example, according to the (1) type, the value ["electron" / probability / coincidence / 0.5 and / a

"network" / 0.3 and / "reception" / 0.2 and / "mail"] 0.3 is obtained to the keyword "e-mail." On the other hand, in this example 1, actual search results are taken in and learned and coincidence probability is amended. For example, when the search formula which the retrieving person gave is "multimedia ORmulti-media", "multimedia" is set to Ki and "multi-media" is set to Kj. new coincidence probability Wij* comes to be shown in the following formula.
The number of times to which one of the number of times/Ki in which Kj carried out OR combination to Wij*= Wij+Ki, and the Kj(s) appeared x (1-Wij) -- (2)
It is shown that 1-Wij of the above-mentioned formula is calculation of the increment of the probability except Wij.
[0019]Thus, study amendment of coincidence probability will increase the similarity between each word frequently used by OR combination. It is stored in the coincidence probability learning data file 11 whenever the basic data for performing such calculation is search (Step S2 of drawing 2). The form of this data file should just list the fact the keyword what kind of keyword carried out OR combination, and was searched, for example. The similarity calculation module 9 calculates (1) type and (2) types to predetermined timing.
[0020]In this way, if the coincidence probability learning data file 11 is generated, the registrant 7 will supply the new keyword for registration (Step S3). The candidate's output is required as the registration interface 8 displaying the high keyword list of the keyword and similarity to the similarity calculation module 9 (step S4). After the similarity calculation module 9 performs similarity calculation with reference to the coincidence probability learning data file 11 (Step S5), it is arranged sequentially from the high thing of similarity, and outputs a candidate (Step S6).
[0021]"mail" is displayed for the coincidence probability 0.6 and an "electron" to the keyword [list / the] for example "e-mail", in the form which the coincidence probability 0.5 and a "network" called the coincidence probability 0.35, "reception" called the coincidence probability 0.2, and "e-mail" called coincidence probability 0.1. Here, the registrant 7 chooses these [all], when registering all the keywords into the synonym dictionary 5. It chooses the part, in registering only a part. Attached information is attached to these keywords, respectively. This attached information is the information that it is a narrower term of the keyword which is the target of registration, for example, and information that a meaning becomes the same when used in what kind of the context. For example, as attached information, a "kind" performs an "electron", as for a "network", an "element" is performed, and, as for "reception", registration which called "operation" and "mail" "English" and "e-mail" called the "kind" as attached information as attached information as attached information is performed as attached information.
[0022]<Effect of the example 1> Since the information which is helpful when judging the similarity between the keywords actually used for search on the basis of coincidence probability is beforehand learned and accumulated in the coincidence probability learning data file 11 as mentioned above, The list of high keywords of similarity is automatically displayed using this, and dictionary registration becomes possible by choosing these. For this reason, a registrant's burden is eased and improvement in the fuzzy retrieval precision at the time of performing a full-text search etc. can be aimed at. In this way, for example, the various data having contained huge new words and technical terms, such as network news, can be flexibly searched now easily. And since it is learned automatically and accumulated also about the term showing the new fact generated every day in order to use actual search results, the burden into which a registrant prepares this kind of word for beforehand, and registers it is mitigable.
[0023]<Example 2> Drawing 3 is an explanatory view of the example 2 of operation. Coincidence probability expressed the similarity between words in the above-mentioned example 1. On the other hand, when searching, AND combination may be carried out and two or more keywords may be searched. Such a keyword adjoins simultaneously in a mutually applicable document, and appears. Therefore, the probability that another side is simultaneously included in the document containing one side is high. Then, these are registered as a high keyword of similarity. Thus, the probability that the word searched by carrying out AND combination mutually will appear is called connection probability. This connection probability is expressed with the following formula.
Connection probability Wij = document number in which one of the number of times / word "i" which the word "i", and the word "j" connected, and the words "j" appeared -- (3)
[0024]For example, in 0.5 and "network" connection probability, "electronic" connection probability will be in the state where 0.3 and the connection probability of "reception" called it 0.2,

to the keyword "e-mail." Here, the result actually used by search by AND combination like the example 1 is accumulated in the connection probability learning data file 12 shown in drawing 3. By this, study amends connection probability like the example 1. The data processing is performed as follows. The form of this formula is the same as the example 1.

The number of times to which one of the number of times/Ki in which Kj carried out AND combination to Wij*= Wij+Ki, and the Kj(s) appeared x (1-Wij) -- (4)

By this, since the similarity between the keywords often searched with AND combination becomes large gradually by study, the accuracy of search results improves in fuzzy search. Usually, when performing retrieval by keyword, most AND combination or OR combination are used. There is the feature of saying in many cases that especially the keyword with high connection probability is connected mutually, and makes an idiom and a compound.

[0025]Also in this example 2, if the search information 10 is inputted in Step S1 as shown in drawing 3, that result will be accumulated in the connection probability learning data file 12 via the similarity calculation module 9 (Step S2). And the registrant 7 supplies a keyword for the registration which is a keyword (Step S3), and the registration interface 8 requires a candidate of the similarity calculation module 9 (step S4). By referring to the connection probability learning data file 12, the similarity calculation module 9 obtains a candidate and outputs this to the registration interface 8 (Step S5, S6). The result is displayed to the registrant 7 (Step S7), and the selected synonym is registered into the synonym dictionary 5 (Step S8).

[0026]For example, to a keyword "e-mail", an "electron" serves as the connection probability 0.7, "reception" serves as the connection probability 0.4, and "software" serves as connection probability 0.2. Therefore, if these [all] are registered, for example, the information which "reception" called the "electron" the "kind", called it "operation" and called "software" the "kind" as the attached information will be doubled and registered, respectively.

[0027]<Effect of the example 2> Since the operating experience about a keyword with high connection probability is stored in the connection probability learning data file 12 as it is from search information and synonym registration can be performed as mentioned above using this according to the example 2, registering operation of the synonym dictionary 5 is made easy. And the accuracy of fuzzy search can be raised like the example 1, and retrieval precision can be raised.

[0028]<Example 3> The explanatory view of the example 3 of operation is shown in drawing 4. This example shows the registration method of the synonym dictionary which mainly took the spelling error of the keyword, etc. into consideration. For example, Mailer "eudora" has. When searching the document containing this keyword, a retrieving person may input "eudra." For a Japanese, an English word has many spelling errors called the substitution of vowel addition, omission, and a consonant. This example is an example of which the vowel "o" dropped out, and it turns out that the whole keyword input by such a spelling error is performed about twenty percent. What replaces r and l accidentally and produces a spelling error is called substitution. There is also an error which adds a hyphen or drops out.

[0029]In order to relieve such a spelling error automatically, adoption of a synonym dictionary which is searched also including the high word of notational similarity is preferred. In this example 3, accumulate the spelling error which is actually easy to produce in the addition omission substitution list file 13, it is made to learn using actual search information, and it registers with the synonym dictionary 5 by making this into a synonym. That is, as a spelling error is also registered as a synonym, it aims at improvement in retrieval precision. Such similarity is called notation similarity. Urban area distance calculation is known by the calculation method of notation similarity, for example.

[0030]The urban area calculation method explanatory view of "eudora" and "eudra" is shown in drawing 5. Here, the footprint is advanced to the upper right corner from the lower left corner by the method of arranging the word of a comparison object one character at a time scatteringly on a vertical axis and a horizontal axis, progressing to a lattice point in the case of the same character, and progressing in the direction vertical in omission etc., or horizontal. In this case, the sum of the number of characters of a vertical axis and a horizontal axis is made into a denominator, the number of the numbers of characters which dropped out is set to 1, and the distance of both keywords is calculated like 1/11. It asks for similarity as 1- (1/11).

[0031]Restriction of the path of dynamic programming utilization time is shown in drawing 6. If

dynamic programming is used when comparing the character of a vertical axis like drawing 5, and a horizontal axis, the path will be restricted to the direction of the right, vertical above, and an inclined upper right direction. It passes along either of these paths, and the shortest path is followed from a lower left corner to an upper right corner. In this way, it asks for the notation similarity of both keywords like 10/11. Based on the similarity calculation result obtained by such a method, a synonymous candidate is outputted to the registration interface 8 sequentially from the high thing of notation similarity. The registrant 7 chooses a suitable thing from among these, and registers with the synonym dictionary 5.

[0032]Other processings are the same as that of the example 1 and the example 2 which it explained until now. For example, notation similarity drops [an "interface"] to 14/15 to a keyword "interface." The attached information serves as contents of the "notation", for example. In this way, by registering, in the case of search, the word from which such spelling differs is also automatically chosen as a keyword, and can raise retrieval precision. And if calculation of notation similarity, etc. are performed each time in the case of search, search time will start for a long time. On the other hand, by choosing the keyword which the spelling error using the synonym dictionary 5 also took into consideration beforehand, the high-speed search techniques, such as dichotomizing search, can be used, and search time can be carried out in a short time.

[0033]In order to calculate notation similarity at high speed, the following techniques are also employable, for example. The matching diaphragm explanatory view of urban area distance calculation is shown in drawing 7. As shown in this figure, the shortest route in the case of calculating similarity using dynamic programming about the arbitrary words "i and j" shown on the vertical axis and the horizontal axis should be included within the range surrounded by this rhombus. Therefore, if the calculation about portions other than this is excepted, the object of calculation is reduced and computation time can be shortened.

[0034]<Effect of the example 3> If the spelling error etc. which a retrieving person tends to start are learned based on actual search results, and are accumulated beforehand as mentioned above and it is made to perform synonym dictionary registration using it, It becomes possible by correcting this automatically, when a user performs a spelling error, or also taking the spelling error of the data itself into consideration, and making the word into a search term to raise retrieval precision more.

[Translation done.]

* NOTICES *

DESCRIPTION OF DRAWINGS

[Brief Description of the Drawings]
[Drawing 1]It is a system block figure for operation of this invention.
[Drawing 2]It is an explanatory view of the example 1 of operation.
[Drawing 3]It is an explanatory view of the example 2 of operation.
[Drawing 4]It is an explanatory view of the example 3 of operation.
[Drawing 5]It is an urban area distance calculation explanatory view.
[Drawing 6]It is a restriction explanatory view of the path of dynamic programming utilization time.
[Drawing 7]It is a matching diaphragm explanatory view of urban area distance calculation.
[Description of Notations]
1 Network
2-1-2-3 Database
3 The terminal for search
4 The terminal for synonym dictionary registration
5 Synonym dictionary
6 Study data file

[Translation done.]